DOT/FAA/AM-01/10

Office of Aerospace Medicine
Washington, DC 20591

Investigating the Validity of Performance and Objective Workload Evaluation Research (POWER)

Carol A. Manning
Scott H. Mills
Cynthia Fox
Elaine Pfleiderer
Civil Aerospace Medical Institute
Federal Aviation Administration
Oklahoma City, Oklahoma 73125
Henry J. Mogilka
FAA Academy
Oklahoma City, Oklahoma 73125

July 2001

Final Report

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

U.S. Department of Transportation
Federal Aviation Administration

NOTICE

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-01/10		

4. Title and Subtitle	5. Report Date
Investigating the Validity of Performance and Objective Workload Evaluation Research (POWER)	July 2001
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Manning, C.A.[1], Mills, S.H.[2], Fox, C.[1], Pfleiderer, E.[1], and Mogilka, H.J.[3]	

9. Performing Organization Name and Address			10. Work Unit No. (TRAIS)
[1]FAA Civil Aerospace Medical Institute P.O. Box 25082 Oklahoma City, OK 73125	[2]SBC Technology Resources, Inc. 9505 Arboretum Blvd. Austin, TX 78759	[3]FAA Academy, Air Traffic Division P.O. Box 25082 Oklahoma City, OK 73125	11. Contract or Grant No.

12. Sponsoring Agency name and Address	13. Type of Report and Period Covered
Office of Aerospace Medicine Federal Aviation Administration 800 Independence Ave., S. W. Washington, D.C. 20591	14. Sponsoring Agency Code

15. Supplemental Notes

Work was accomplished under approved subtask AM-B-00-HRR-516.

16. Abstract

Performance and Objective Workload Evaluation Research (POWER) software was developed to provide objective measures of ATC taskload and performance. POWER uses data extracted from National Airspace System (NAS) System Analysis Recording (SAR) files to compute a set of objective measures. A study was conducted to investigate the relationship of POWER measures with measures of sector complexity, controller workload, and performance. Sixteen instructors from the FAA Academy in Oklahoma City, OK, watched eight traffic samples from four en route sectors in the Kansas City Center using the Systematic Air Traffic Operations Research Initiative (SATORI) system. POWER measures were computed using the same data. Participants made three estimates of the workload experienced by radar controllers and provided two types of assessments of their performance. Sector complexity was determined using information about sector characteristics and the traffic samples. Some POWER measures were related to sector complexity and controller workload, but the relationship with performance was less clear. While this exploratory study provides important information about the POWER measures, additional research is needed to better understand these relationships. When the properties and limitations of these measures are better understood, they may then be used to calculate baseline measures for the current National Airspace System.

17. Key Words	18. Distribution Statement
Air Traffic Control, Workload, Taskload, Performance, Sector Complexity	Document is available to the public through the National Technical Information Service, Springfield, Virginia, 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	40	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Investigating the Validity of Performance and Objective Workload Evaluation Research (POWER)

Introduction

Need for Measuring ATC Workload, Taskload, Complexity, and Performance

To understand how new air traffic control (ATC) systems and procedures may affect individual air traffic controllers and the ATC system as a whole, it is necessary to measure the inter-relationships of mental workload, taskload, sector complexity, and controller performance in ATC (Wickens, Mavor, Parasuraman, & McGee, 1998). The effects of using different display designs or alternative procedures on controllers' workload and performance must be determined before they are implemented. When new ATC systems are introduced in field facilities, it is necessary to document their effects on individual and system performance, both soon after implementation and later, after controllers have become accustomed to using them. Computing measures of taskload and performance on a system level, while accounting for sector complexity, may also contribute to better prediction of overloads at specific sectors.

Defining Controller Workload, Taskload, Sector Complexity, and Performance

While many methods have been used to measure ATC workload, taskload, sector complexity, and controller performance, definitions of these terms are not widely agreed upon. In general, "workload" typically refers to the physical and mental effort an individual exerts to perform a task. In this sense, ATC workload may be differentiated from "taskload" in that "taskload" refers to air traffic events to which the controller is exposed, whereas "workload" describes the controller's reaction to the events and the perceived effort involved in managing the events.

"Sector complexity" describes the characteristics (both static and dynamic) of the air traffic environment that combine with the taskload to produce a given level of controller workload (Grossberg, 1989). In that sense, "complexity" can mediate the relationship between taskload and workload.

According to Federal Aviation Administration (FAA) Air Traffic Control (Order 7110.65M, 2000) states "The primary purpose of the ATC system is to prevent a collision between aircraft operating in the system and to organize and expedite the flow of traffic." Thus, measurement of controller performance involves determining the effectiveness with which an individual controller's activities accomplish these goals.

Methods for Measuring ATC Workload, Taskload, Sector Complexity, and Performance

Many methods have been developed to measure workload, taskload, sector complexity, and controller performance (see Hadley, Guttman, & Stringer, 1999, for a database containing 162 of these measures). The dynamic nature of ATC (requiring controllers to both predict movements of individual aircraft and evaluate changes in the relative positions of groups of aircraft) makes it necessary to take the passage of time into consideration when measuring these constructs. When time is considered, it is even more difficult to measure controller performance and workload than it is to measure taskload and sector complexity. The reason is that taskload may be measured by counting recorded ATC events, and sector complexity can be measured by recording observable sector characteristics and other observable factors about the ATC situation. Controller workload and performance, on the other hand, include factors that cannot be easily observed and are, therefore, not as easy to measure. For example, controllers continually review aircraft positions, directions, and speeds, and mentally project aircraft positions, but take observable actions less frequently. It is possible to count or otherwise evaluate certain observable activities, such as making keyboard entries and marking or moving flight progress strips. However, the relationship between these measures (taskload) and the amount of cognitive effort expended (mental workload) or the effectiveness of the results (performance) is unclear. Even actions that

appear to be interpretable (e.g., commission of operational errors resulting in losses of separation) may not be very meaningful because they occur so infrequently as to be of little value in assessing individual performance; also because it is often difficult to determine their cause, largely because of the dynamic nature of the task.

Mental Workload Measures

Workload, the controller's cognitive reaction to the taskload experienced, is hypothesized to include components that cannot be easily explained by measuring taskload alone. Because most of a controller's activities are cognitive, not physical, it is more appropriate to measure mental, rather than physical workload. Measures of mental workload in ATC are typically obtained either during a simulated scenario or after its completion. One measure, the NASA Task Load Index (TLX; Hart & Staveland, 1988) is given to controllers after they finish a scenario. To complete the NASA TLX, controllers provide separate ratings for each of six scales: Mental demand, physical demand, temporal demand, effort, frustration, and performance.

In contrast, the Air Traffic Workload Input Technique (ATWIT) measures mental workload in "real-time" (Stein, 1985). The ATWIT presents auditory and visual cues (a tone and illumination, respectively) that prompt a controller to press one of seven buttons within a specified amount of time to indicate the amount of mental workload experienced at that moment. The Workload Assessment Keypad (WAK) device records each rating as well as the time it took to respond to the prompt.

The primary advantage of using a real-time mental workload measure is that the respondent can report the experience while or soon after it occurs. However, requiring a controller to provide a real-time mental workload estimate in addition to the other tasks that must be performed may increase a controller's perceived mental workload or, worse yet, may interfere with the performance of the remaining tasks. On the other hand, obtaining a mental workload rating from a controller after a scenario is complete may be overly influenced by earlier or later events (i.e., primacy or recency effects) and the controller may forget to consider certain events altogether. The unidimensional nature of a real-time workload rating as compared with a group of post-scenario workload ratings based on a set of multi-dimensional rating scales must also be considered.

Taskload Measures

Several measures describing controller taskload have been derived from recordings of either simulation data or operational National Airspace System (NAS) activities. For example, Buckley, DeBaryshe, Hitchner, & Kohn (1983) developed a set of computer-derived measures obtained during ATC simulations. They identified four factors that summarized the measures: conflict, occupancy, communications, and delay. Galushka, Frederick, Mogford, & Krois (1995) used counts of controller activities, as well as Over-the-Shoulder (OTS) subjective performance ratings to assess en route air traffic controller baseline performance during a simulation study.

Using data extracted from the Log and Track files generated by the Data Analysis and Reduction Tool (DART; Federal Aviation Administration, 1993), Mills, Manning, & Pfleiderer (1999) developed an extensive set of computer-derived taskload measures. Performance and Objective Workload Evaluation Research (POWER) extracts recorded information to compute measures such as numbers of controlled aircraft, altitude changes, specific controller data entries and data entry errors, numbers and durations of handoffs, and variations in aircraft headings, speeds, and altitudes. (See Table 1, below, for a complete list of measures.)

Sector Complexity Measures

Several measures of sector complexity have also been developed. Complexity measures typically include physical characteristics of a sector, procedures employed in the sector, and factors related to the specific air traffic situation that may increase its perceived difficulty. For example, Grossberg (1989) identified three groups of factors (control adjustments such as merging, spacing, and speed changes; climbing and descending flight paths; and mix of aircraft types) that contributed to the complexity of operations in different sectors.

Mogford, Murphy, Roske-Hofstrand, Yastrop, & Guttman (1994) used multidimensional scaling techniques to identify 15 complexity factors. These were 1) number of climbing or descending aircraft, 2)

degree of aircraft mix, 3) number of intersecting aircraft flight paths, 4) number of multiple functions the controller must perform, 5) number of required procedures to be performed, 6) number of military flights, 7) coordination with other sectors or facilities, 8) extent to which hubbing is a factor, 9) extent to which weather affects ATC operations, 10) number of complex aircraft routings, 11) special-use airspace, 12) size of sector airspace, 13) requirement for longitudinal sequencing and spacing, 14) adequacy of radio and radar coverage, and 15) amount of radio frequency congestion.

Wyndemere, Inc. (1996) identified 19 factors that they believed contributed to complexity in air traffic control. These were 1) airspace structure, 2) special use airspace, 3) weather effects on airspace structure, 4) proximity of potential conflicts to sector boundary, 5) aircraft density, 6) number of facilities served by a sector, 7) number of aircraft climbing or descending, 8) number of crossing altitude profiles, 9) weather effects on aircraft density, 10) variance in aircraft speed, 11) variance in directions of flight, 12) performance mix of aircraft, 13) winds, 14) distribution of closest points of approach, 15) angle of convergence in conflict situation, 16) neighbors (proximity of aircraft pairs), 17) level of knowledge of aircraft intent, 18) separation requirements, and 19) coordination.

Although some of the specific complexity factors proposed by different authors are not identical, the complexity construct has been found useful in research. For example, Rodgers, Mogford, & Mogford (1998) found a significant multiple correlation between the overall rate of operational errors at Atlanta Center and Mogford et al.'s (1994) 15 complexity factors.

While they seem somewhat similar, complexity factors differ from taskload measures. Complexity factors include a number of variables related to a sector's static structure and characteristics, established functions and procedures that apply to a sector, and percentages of aircraft that meet a particular criterion. On the other hand, taskload measures are statistics that describe distributions of controller and aircraft activities.

If information about sector and traffic characteristics is available, it should be relatively easy to derive values for most of the sector complexity measures. Though the constructs proposed by different authors are closely related, unfortunately, the number of factors necessary to describe sector complexity remains unclear. Nevertheless, it appears that the complexity construct may provide information beyond what is available from the measurement of taskload alone.

Controller Performance Measures

Subject Matter Expert (SME) observations. One of the challenges associated with measuring controller performance is evaluating the different approaches controllers use to control traffic. Most techniques a controller may use to successfully maintain aircraft separation and a smooth flow of air traffic are considered acceptable. However, such individual techniques make it difficult to evaluate the effectiveness of an individual controller's actions to move a set of aircraft through a sector. To accommodate differences in technique, SME observations are often used to measure controller performance.

Several procedures have been developed to record SME observations of controller performance. The Behavioral Summary Scales (BSS; Borman et al., 2001) were developed as a criterion measure against which the Air Traffic Selection and Training (AT-SAT) selection battery (Ramos, Heil, & Manning, 2001) could be validated. The BSS scales included ten distinct performance categories and measured "typical" rather than "maximum" performance; that is, how well controllers performed consistently over time, rather than how well they could perform under peak traffic conditions.

Several other procedures have been developed to evaluate controller performance during "maximum" conditions (during difficult high-fidelity simulations). For example, Bruskiewicz, Hedge, Manning, & Mogilka (2000) developed two procedures for measuring controller performance to use in a high-fidelity simulation study conducted to evaluate the AT-SAT criterion performance measures. These were an Over-the-Shoulder (OTS) rating form and a Behavior and Event Checklist (BEC). The OTS rating form, used to evaluate controller performance across broad dimensions, was based in part on the BSS. The BEC was used to record specific mistakes made during the simulation exercises.

The advantage of using SME observations as a basis for evaluating controller performance is that SMEs (especially instructors involved in controller training) possess detailed knowledge about the job and, thus, can evaluate aspects of controllers' behavior beyond what can be obtained from merely counting

events. Many SMEs are also very accustomed to observing and evaluating the actions of other controllers to provide feedback for trainees.

However, several problems may be associated with SME observations. First, determining appropriate performance ratings and identifying mistakes requires the observer to make extensive interpretations. To assure the reliability of these subjective ratings and error counts, extensive SME training and practice sessions are required. Even when they are trained, it is difficult to determine whether SMEs are focusing on cues relevant to performance. It is also not always possible to obtain SME observations because quite often, an insufficient number of controllers is available to participate in these activities.

Controller-generated responses on dynamic tests. Another way to measure controller performance is to have controllers answer questions that test their job knowledge or judgment in a dynamic way. The Controller Decision Evaluation (CODE) technique, developed by Buckley & Beebe (1972), presented controllers with filmed simulations of air traffic situations and asked them to answer related questions. A simplified version of the CODE, the Multiplex Controller Aptitude Test (MCAT; Dailey & Pickrel, 1984), was developed for ATC job applicants. The idea of using a dynamically administered written test carried over into the ATC training environment. Controller Skills Tests (CSTs; Tucker, 1984) were developed to test students in the ATC screening programs by requiring them to quickly interpret air traffic information and then answer multiple choice questions.

Hanson et al. (1999) developed a Computer Based Performance Measure (CBPM) to provide another criterion measure for the AT-SAT project. The CBPM presented dynamic ATC situations (including simulated voice communications) and asked controllers to answer a series of multiple-choice questions to identify potential conflicts, sequence aircraft, and demonstrate control judgment.

Tests requiring controllers to choose between responses are desirable performance measures, from the researchers' point of view, because they produce easily scored responses that can be summed into test scores and directly compared with test scores earned by other controllers. However, to consider these scores to be valid ATC performance measures, it must first be assumed that controllers who accurately choose a response when observing an air traffic situation can also perform ATC tasks effectively. Although scores on the CBPM were positively correlated with both scores on the BSS (indicating typical performance) and measures used to evaluate performance in the AT-SAT high-fidelity simulations (indicating maximum performance), the correlations were not sufficiently high to eliminate the ambiguity in their interpretation. Thus, the appropriateness of equating performance on dynamic, multiple-choice tests with performance in controlling traffic may still be questioned.

Purpose of Study

Our challenge was to develop a set of measures describing different aspects of ATC activity that are objective, reliable, valid, and relatively easy to obtain. It is desirable to use routinely recorded data because SME observations and mental workload ratings, which may have more "face validity" than taskload measures, may be influenced by rater biases and are often not available. Recorded ATC data are not subject to the same rater biases and usually are available. On the other hand, it is possible that taskload measures are not adequate if numbers derived from recorded ATC data do not sufficiently account for subtle aspects of controller workload and performance.

This study utilized the POWER measures described above to measure controller taskload. POWER measures encompass counts of aircraft and controller activities computed from routinely-recorded ATC data. While a set of measures has been derived in POWER, as yet, no empirical evidence is available to indicate whether these numbers actually measure the constructs they were intended to measure or how stable that relationship might be. For example, while we might generally predict that a controller who takes more actions is less efficient, such a relationship may not be invariant but may, instead, be influenced by external factors such as weather or sector complexity. To begin to answer these questions, this study was conducted to examine the relationship between the POWER measures, SME ratings of mental workload and controller performance, and measures of sector complexity.

In particular, we predicted that some POWER measures may be related to measures of sector complexity, some may be related to controller performance, and some may be related to mental workload (See Table 1). If some of the POWER measures are related to measures of sector complexity, mental

workload, and/or controller performance, it may be possible to use them in situations where it would not otherwise be possible to evaluate these variables (when SMEs are unavailable or controllers are unable to provide workload evaluations). For example, a validated set of POWER measures could provide information that would allow post-implementation evaluation of the operational effects of new ATC systems using routinely recorded ATC system data.

To assess the relationships between POWER measures and measures of sector complexity, mental workload, and controller performance, a set of POWER measures derived from recorded ATC data was compared with SME-derived mental workload and controller performance measures obtained from the same data source. While mental workload ratings are usually obtained from the specialists who controlled the traffic being analyzed, and controller performance ratings are usually obtained from direct SME observations of controllers' performance, only recorded ATC data were available for this study. Therefore, rather than observing and rating controller performance as it occurred, SMEs who participated in the study evaluated controller performance by observing the re-creations of available recorded data. In addition, the SMEs in the study rated the mental workload they inferred occurred during the observations, instead of having controllers rate their own mental workload. The use of this methodology may be criticized because subjective workload esti-

Table 1. *Expected Relationships Between POWER Measures and Measures of Sector Complexity, Controller Performance, and Subjective Workload.*

	Expected Relationships		
Power Measure	Sector Complexity	Controller Performance	Subjective Workload
Total N aircraft controlled	X		X
Max aircraft controlled simultaneously	X		X
Average time aircraft under control	X	X	X
Avg Heading variation	X	X	X
Avg Speed variation	X	X	X
Avg Altitude variation	X	X	X
Total N altitude changes	X	X	X
Total N handoffs	X		X
Total N handoffs accepted			X
Avg time to accept handoff		X	X
Total N handoffs initiated			X
Avg time until initiated HOs are accepted			X
N Radar controller data entries		X	X
N Radar controller data entry errors		X	X
N Data controller data entries			X
N Data controller data entry errors			X
N Route displays		X	X
N Radar controller pointouts		X	X
N Data controller pointouts			X
N data block offsets		X	X
Total N Conflict Alerts Displayed		X	X
Number of Conflict Alert suppression entries		X	
N Distance Reference Indicators requested		X	X
N Distance Reference Indicators deleted		X	X
N track reroutes			X
N strip requests			X

mates typically depend in part on the subject's individual differences in skill and stress tolerance. Asking observers removed from the observations to infer the workload experienced by someone they cannot see may wash out some of the variability in the workload estimates. On the other hand, sufficient cues about the controller's reaction to the situation may be available for the trained SMEs to reliably determine how the controller is handling the taskload present during traffic samples.

Method

Participants

Participants were 16 en route air traffic control instructors from the FAA Academy in Oklahoma City, OK. All had previously served as fully-qualified controllers at en route Air Route Traffic Control Centers (ARTCCs.) Two participants had controlled traffic at some of the sectors represented in the traffic samples, though none had worked at all the sectors included in the study. All participants were fully-qualified instructors who had received training on methods for observing and evaluating controller performance.

Materials

Traffic Samples

System Analysis Report (SAR) and voice communication tapes were obtained for 12 traffic samples recorded during January 1999, at four ATC sectors in the Kansas City ARTCC. The traffic samples consisted of routine operations and contained no accidents or incidents. The SAR data used for the traffic samples were extracted by DART and the National Track Analysis Program (NTAP; Federal Aviation Administration, 1991). Resulting files were processed both by Systematic Air Traffic Operations Research Initiative (SATORI; Rodgers & Duke, 1993) and POWER software (Mills, Manning, & Pfleiderer, 1999). SATORI synchronizes information from DART and NTAP files with tapes containing the Radar (R) controller's voice communications, using the time code common to both data sources, while POWER uses a subset of the DART files to compute measures of sector and controller activity.

Three traffic samples were re-created for each of the four sectors. One traffic sample for each sector (used for training) was eight minutes long. The two remaining experimental traffic samples for each sector were both 20 minutes long.

Sector Training Materials

Computerized training sessions were shown to participants that described the characteristics and procedures applicable to each sector. Participants also examined copies of sector maps on which important sector information was highlighted. These maps and a copy of the sector binder (containing additional sector information) were available for the participants to review while they watched the traffic samples. Participants also had access to flight plan information (derived from flight strip messages) for each aircraft controlled by the sector during the traffic sample.

Mental Workload Measures

Participants provided three types of measures describing the mental workload they thought the R controller experienced during each traffic sample. The ATWIT presented a tone and illumination that prompted the participant to press one of seven buttons within a 20-second period. In this study, ATWIT ratings were collected every four minutes during each traffic sample using the Workload Assessment Keypad (WAK; see Appendix A). Participants were instructed to enter ATWIT ratings that indicated the amount of mental workload they thought the R controller experienced in reaction to the taskload that occurred during the traffic sample.

The second type of mental workload measure was a modified version of the NASA Task Load Index (TLX; Hart & Staveland, 1988). TLX ratings were obtained after each traffic sample had ended. Separate ratings were provided for each of the six TLX scales. Participants were instructed to base their TLX ratings on how difficult they thought the R controller's task was and how well they thought the R controller controlled the traffic. The TLX ratings were entered using a computerized screen that allowed ratings to be changed before they were finalized (See Appendix B). Participants only provided ratings on the individual scales but did not perform the associated dimensional weighting procedure because 1) previous research suggests that there is little difference in

the result produced by using the weighted and unweighted composites (see Moroney, Biers, & Eggemeier, 1995) and 2) the process used to obtain the weights is "ineffective" (Nygren, 1991).

Instructions for completing the TLX are shown in Appendix C. Note that the TLX scales were labeled "Low" (on the left side of the scale) and "High" (on the right side of the scale) for all scales except TLX Performance, for which the left side was labeled "Good" and the right side was labeled "Poor." A zero was assigned to the left-most rating, while 100 was assigned to the right-most rating on each scale. Thus, a lower rating on the numerical TLX Performance scale corresponded with better performance.

The third type of mental workload measure was a rating of the activity level the participant perceived to occur during each traffic sample. The activity level rating used a 5-point scale ranging from "Not at all busy" to "Very busy" (see Appendix D.) The activity level rating was provided after the completion of each traffic sample.

Controller Performance Measures

Two controller performance measures were used in this study. Both were based on measures previously developed for the AT-SAT high-fidelity simulation study (Bruskiewicz, Hedge, Manning, & Mogilka, 2000). The first, the Over-the-Shoulder (OTS) rating form, was used to evaluate controller performance across broad rating dimensions. In this study, participants used a revised version of the OTS form originally developed for the AT-SAT high-fidelity validation study. Unlike the raters who observed controller performance during the AT-SAT high-fidelity validation study, the participants in this study had access to only the R controllers' voice communications (even when a Data [D] controller was also working) and, thus, may have been unable to evaluate all of the events that occurred at the sector during the traffic sample. For example, the Coordinating, Performing Multiple Tasks, and Managing Sector Workload rating dimensions from the original version of the OTS form were removed from this version because D controller communications were unavailable.

Lack of availability of other information further reduced the number of rating dimensions that could be used for this version of the form. For example, the rating dimension "Maintaining Attention and Situation Awareness" from the original form included the behavioral example, "data block overlap." In this study, it was not possible for the participant to determine whether data blocks actually overlapped during the traffic sample because 1) the size of the display used to present the traffic samples was not the same as the size of the display the controller originally used when controlling traffic, and 2) the length of the leader line separating the target from the data block was not known (because the analog switch used to set the length was not recorded). Because participants could not determine whether or not data blocks actually overlapped, they were not able to effectively evaluate whether the controller maintained attention and situation awareness, and so that rating dimension was eliminated from the form.

The resulting set of rating dimensions included on the POWER OTS rating form included: Maintaining Separation; Maintaining Efficient Air Traffic Flow; Communicating Clearly, Accurately, and Efficiently; Technical Knowledge; Prioritizing, and Overall Effectiveness (see Appendix E for a copy of the form). Each rating dimension included several behavioral examples that participants could review when completing the form. Instructions for using the POWER OTS rating form to evaluate a controller's performance (based on recorded traffic samples) are shown in Appendix F.

The second controller performance measure used in the study was the Behavior and Event Checklist (BEC; see Appendix G). Participants used the BEC to record mistakes they determined that the R controller made during the traffic sample. The error categories on the BEC were Operational Errors (OEs), Operational Deviations (ODs)/Special Use Airspace (SUA) Violations, Fail to Accept Handoff, Letter of Agreement (LOA)/Directive Violations, Transmission Errors, Made Late Frequency Change, Unnecessary Delays, Incorrect Information in the Computer, and Fail to Issue Weather Information. Instructions describing how participants should identify the errors listed on the BEC are shown in Appendix H.

Sector Complexity Measures

The sector complexity measures used in this study were based on Mogford et al.'s (1994) 15 complexity factors. Mogford's factors were combined into two complexity measures: static complexity and dynamic complexity. The static complexity measure included variables that remained constant over the course of a traffic sample. These were airspace size and the num-

bers of: 1) sectors adjacent to the controlling sector, 2) transfer control points in the sector, 3) sequencing functions utilized in the sector, 4) military operations, 5) major airports in the sector, 6) VORTACS, 7) airway and jetway intersections, 8) miles of airways, and 9) shelves. This information was derived from sector descriptions available in the sector binder, letters of agreement for each sector, and Kansas City ARTCC's Adaptation Control Environmental System (ACES) map files.

The dynamic complexity measure included variables related to each sector that would be expected to vary during a traffic sample. These were numbers of 1) pilot/controller transmissions, 2) interphone communications (with another controller), 3) military aircraft, 4) heading changes or vectors issued, 5) altitude and speed restrictions issued, 6) conversations about holding, and 7) conversations about weather. Also included were maximum Hs and Ls displayed during a traffic sample (indicating high and low weather activity), amount of climbing/descending traffic, percentages of jets and VFR aircraft controlled during the traffic sample, percentages of arrivals/departures for the St. Louis airport, and a variable reflecting traffic volume (amount of traffic per volume of airspace). This information was derived from the traffic samples.

The static and dynamic complexity measures were then computed by averaging standardized scores for each of the corresponding variables. An overall complexity measure was also computed by averaging standardized scores for all variables included in either the static or dynamic complexity measures.

Procedure

Participants read a description of the purpose and method of the experiment, completed consent and biographical information forms, then reviewed the instructions for completing the workload and performance measures. For each of the four sectors, participants then a) reviewed sector-specific training materials, b) observed one 8-minute training traffic sample, and c) observed two 20-minute experimental traffic samples. To ensure continuity, all traffic samples for a sector were shown together as a block. The order in which the four blocks of traffic samples were observed was counter-balanced, as was the order of presentation of the two experimental traffic samples within each block.

While watching each traffic sample, participants used the BEC to record any mistakes they observed. The ATWIT aural signal occurred every four minutes. Participants responded by entering a number between 1 and 7 on the WAK keypad. After each traffic sample was stopped, participants completed the computerized version of the NASA TLX, summed the errors they had marked on the BEC, then completed the OTS rating form. Finally, they rated the activity level for that traffic sample.

Reviewing the training materials and observing the three traffic samples for each sector required about 1½ hours. After observing the traffic samples for all four sectors, participants answered questions about their experiences during the observation process.

Results

The results are presented in two parts. Part one presents findings related to the mental workload and controller performance measures obtained from SMEs observing recorded ATC activities. The analyses a) examined the reliability of the measures, b) described the relationships among the measures, and c) identified a smaller number of measures that can be used to explain most of the variance in the complete set of mental workload and controller performance measures. Part two presents the findings related to how sector complexity, controller performance, and mental workload measures relate to the POWER measures.

Part 1: Analysis of Mental Workload and Controller Performance Measures

Characteristics of the individual mental workload and controller performance measures were examined first. Table 2 shows descriptive statistics for the three mental workload measures (the six TLX scales, the average ATWIT rating, and the SME activity level rating), and the two controller performance measures (the six OTS rating dimensions and the ten BEC items). Mean TLX ratings were low for the workload-related scales and were slightly above the midpoint for the (reverse scaled) Performance rating. All average OTS ratings were slightly below the midpoint. Average counts for most errors were typically low, with Transmission errors being marked most frequently. Standard deviations for the error counts were fairly high in comparison with the means, indicating a lack of agreement between raters. Specifically,

Table 2. *Means and Standard Deviations for Individual Items from Workload and Performance Scales Averaged Over All Traffic Samples (N=128).*

Measure	Mean	Standard Deviation
Mental workload		
TLX Scales (0-100)		
Mental Demand	35.90	20.46
Physical Demand	31.80	19.67
Temporal Demand	33.28	19.58
Performance (100-0)	44.53	16.07
Effort	31.77	18.22
Frustration	25.44	22.32
Average ATWIT rating (1-7)	2.76	1.00
SME activity level rating (1-5)	2.26	.90
Controller performance		
OTS Ratings (1-7)		
Maintaining Separation	3.89	1.01
Maintaining Efficient Air Traffic Flow	3.94	0.91
Communicating Clearly, Accurately, & Efficiently	3.48	1.08
Technical Knowledge	3.88	1.03
Prioritizing	3.75	1.24
Overall Effectiveness	3.72	0.98
BEC Counts		
Operational Errors	.03	.25
Operational Deviations/ SUA violations	.38	.98
LOA/Directive violations	1.0	1.39
Transmission errors	2.02	2.37
Failed to accommodate pilot request	.40	.79
Failed to accept handoff	.05	.23
Made late frequency change	.36	.66
Unnecessary delays	.59	.94
Incorrect information in computer	.80	1.51
Failed to issue weather information	.80	1.29

in most of the traffic samples, no errors of any kind were recorded. However, occasionally, a few observers recorded one or more errors. So for example, although the traffic samples did not include any officially-designated operational errors, two participants thought they observed two operational errors. Thus, the mean OE count was near (but not exactly) zero, while the standard deviation was much higher than the mean.

Analysis of Mental Workload Measures

Interrelationships among the three types of mental workload measures were examined next. Before analyzing the data, an analysis was conducted to assess the reliability of the participants' responses. For the six TLX scales, the ATWIT, and the activity level ratings, the average measure intraclass correlation for the participants was .98. Thus, all participants' data were retained for further analysis.

Table 3 shows intercorrelations among the mental workload measures. The Mental, Physical and Temporal Demand scales, and the Effort scale were all highly correlated (r = .85 or above). The Frustration scale was also significantly correlated with the other TLX scales, but these correlations were not as high. For example, none of the correlations of Frustration with Mental, Physical, and Temporal Demand and Effort exceeded .60. Likewise, the correlation between Frustration and Performance was .24, which, while statistically significant, accounted for just over 5% of the variance. Frustration was the only TLX scale that correlated with Performance.

The Activity Level scale was highly correlated with the ATWIT (r = .84). ATWIT and the Activity Level measure had similar patterns and magnitudes of correlations with the other mental workload variables. Approximate correlations of both these variables with Mental, Physical, and Temporal Demand were .70, with Effort were .60, with Frustration were .40, and with Performance, .10 or less.

Because the TLX Performance scale had such low correlations with most of the other mental workload measures, it was analyzed with the controller performance measures. A principal components analysis was then conducted to derive a reduced set of components that could be used to describe the variance in the remaining mental workload measures. Two factors were derived from this analysis. Table 4 shows eigenvalues and the percent of variance accounted for by the solution. Although the eigenvalue for the second factor was less than 1, it accounted for 12% of the variance in the mental workload measures, so a 2-factor solution was chosen.

Table 3. *Intercorrelations of Mental Workload Measures (N=128).*

	Mental Demand	Physical Demand	Temporal Demand	Performance	Effort	Frustration	ATWIT	Act Lvl
Mental Demand	1.0							
Physical Demand	.94**	1.0						
Temporal Demand	.95**	.92**	1.0					
Performance	-.07	-.06	-.09	1.0				
Effort	.88**	.86**	.89**	-.04	1.0			
Frustration	.56**	.58**	.60**	.24**	.59**	1.0		
ATWIT	.72**	.71**	.73**	-.03	.64**	.32**	1.0	
Act Lvl	.71**	.70**	.70**	-.09	.63**	.36**	.84**	1.0

Note: ** $p < .01$; Act Lvl = Activity Level.

Table 4. *Eigenvalues for Principal Components Analysis of Mental Workload Measures.*

Component	Eigenvalue	% of Variance
1	5.30	75.8
2	0.84	12.0
3	0.42	6.0

Table 5. *Varimax-rotated Component Matrix for Mental Workload Measures.*

Workload Measure	Component 1 - Activity	Component 2 - Frustration
Mental Demand	**.70**	**.65**
Physical Demand	**.68**	**.66**
Temporal Demand	**.68**	**.69**
Effort	**.59**	**.71**
Frustration	.06	**.91**
ATWIT rating	**.92**	.19
Traffic Sample Activity Level	**.90**	.21

Note: Correlations greater than .30 are bolded.

Table 5 shows the varimax-rotated component matrix, which contains correlations of each mental workload measure with the two principal components. The Mental, Physical, and Temporal Demand and Effort scales had correlations of about .6 or higher with both components. Because these scales correlated so highly with both components, their meaning did not contribute significantly to the interpretation of either one. Component 1 was primarily defined by the ATWIT and Activity level ratings so it was labeled "Activity." Component 2 was primarily defined by the TLX Frustration scale, so it was labeled "Frustration."

Analysis of Controller Performance Measures

OTS ratings. Because there were so many individual performance items, the OTS and BEC items were analyzed separately. Before analyzing the OTS data, an assessment of the reliability of participants' ratings was conducted. For the six OTS rating dimensions, the participants' average measure intraclass correlation was .77. Examination of the correlations between ratings revealed that two participants' ratings were negatively correlated with ratings from many of the other participants. When their ratings were removed from the analysis, the resulting average measure intraclass correlation increased to .86. Consequently, OTS ratings for those two participants were excluded from further analysis.

Intercorrelations of the OTS performance measures for the remaining participants are shown in Table 6. Correlations among rating dimensions were statistically significant but not as high as expected, ranging from about .40 to .63. In contrast, Bruskiewicz, Hedge, Manning, & Mogilka (2000) found that the correlations for the seven individual OTS scales used in the AT-SAT high-fidelity simulation study ranged from .80 to .97. We believe that the differences in correlations are probably due to the difference in the amount of time available during the two studies to train the observers.

Because correlations of individual items with the Overall Effectiveness Rating were somewhat low (ranging from .64 to .77), an Average OTS Rating was computed across the five individual OTS scales. Correlations of each individual OTS scale with this Average OTS Rating (shown in the last row of Table 5) were higher than with the Overall Effectiveness Rating, ranging from .72 to .86. A principal components analysis of the OTS ratings produced one component, which correlated .995 with the Average OTS

Table 6. *Intercorrelations of Scales from Over-the-Shoulder Rating Form (N=112).*

	Maintaining Separation	Maintaining Efficient ATC Flow	Communi-cating	Technical Knowledge	Priori-tizing	Overall effectiveness
Maintaining Separation	1.0					
Maintaining Efficient ATC Flow	.52**	1.0				
Communi-cating Clearly, Accurately, Efficiently	.51**	.56**	1.0			
Technical Knowledge	.49**	.60**	.56**	1.0		
Prioritizing	.42**	.61**	.58**	.63**	1.0	
Overall effectiveness	.64**	.68**	.77**	.70**	.66**	1.0
Average OTS rating	.72**	.81**	.80**	.82**	.83**	N/A

Note: * $p < .05$; ** $p < .01$

Rating (and only .91 with Overall Effectiveness). Thus, the Average OTS rating scale was retained as the representative measure describing OTS performance.

BEC items. Before analyzing the BEC data, an analysis was conducted in which the participants were analyzed if they were items to assess the reliability of their error counts. For the ten BEC items, the participants' average measure intraclass correlation was .90. Examination of the correlations between ratings revealed that two participants' ratings were negatively correlated with ratings from some other participants. (These were different participants than those whose OTS ratings were inconsistent with other participants' ratings.) When their ratings were removed from the analysis, the resulting average measure intraclass correlation increased to .92. The small increase in the average measure intraclass correlation resulting from removing the two participants from the analysis did not seem to warrant eliminating their data. Consequently, BEC items for all participants were included in further analyses.

Intercorrelations among items on the Behavior and Event Checklist are shown in Table 7. Recall that these items were counts of different types of errors that participants observed during the traffic sample. Most correlations were moderate in size (the highest was .42), and several were statistically significant. Specifically, the items Failed to Accommodate Pilot Requests, Failed to Accept Handoff, and Failed to Issue Weather Information were significantly correlated with five other variables, and the items Incorrect Information in Computer and LOA/Directive Violations were significantly correlated with four other variables.

A principal components analysis was conducted to summarize the BEC items. Sixteen participants rated eight traffic samples, resulting in 128 evaluations, each consisting of ten BEC items. Three components with eigenvalues greater than 1.0 were extracted that accounted for about 50% of the variance in the data. Table 8 shows the component matrix rotated using the Varimax method. Component loadings greater than or equal to .30 are bolded to highlight items having high relationships with the components.

Component 1 included five of the ten BEC items: Failed to Accept Handoffs, Failed to Issue Weather Information to Pilots, Letter of Agreement or other Facility Directive Violations, Made Late Frequency Changes, and Failed to Accommodate Pilot Requests. Operational Errors also had a positive (though small) correlation with this component. High numbers of these errors occurred during one traffic sample in

Table 7. *Intercorrelations of Items from Behavior and Event Checklist (N=128).*

	OEs	ODs	LOA	Trans	FAPR	FAH	MLFC	UD	IIC	FIWI
Operational Errors (OEs)	1.0									
Operational Deviations/SUA violations (ODs)	.14	1.0								
LOA/Directive violations (LOA)	.14	.08	1.0							
Transmission errors (Trans)	-.05	.19*	.03	1.0						
Failed to accommodate pilot request (FAPR)	.02	.11	.23**	.16	1.0					
Failed to accept handoff (FAH)	.25**	.05	.42**	-.02	.27**	1.0				
Made late frequency change (MLFC)	.03	.03	.13	.07	.19*	.34**	1.0			
Unnecessary delays (UD)	-.01	-.07	.06	.26**	.14	.07	.06	1.0		
Incorrect information in computer (IIC)	.02	.34**	.22**	.16	.19*	.12	.08	.17	1.0	
Failed to issue weather information (FIWI)	.07	.03	.34**	-.05	.26**	.33**	.27**	.03	.25**	1.0

Note: * $p < .05$; ** $p < .01$

Table 8. *Varimax-rotated Component Matrix for BEC Items (N=128).*

	Component		
BEC item	Component 1- Inactivity	Component 2 - Disorganization	Component 3 - Inefficient but safe
Operational Errors	.23	**.31**	**-.46**
Operational Deviations/SUA violations	-.05	**.86**	-.09
Failed to accept handoff	**.77**	.05	-.12
LOA/Directive violations	**.64**	.20	-.08
Transmission errors	-.05	**.37**	**.66**
Failed to accommodate pilot requests	**.51**	.16	**.33**
Made late frequency change	**.57**	-.10	.15
Unnecessary delays	.15	-.02	**.71**
Incorrect information in computer	.23	**.65**	.22
Failed to issue weather information	**.69**	.06	-.03

Note: Correlations greater than .3 are bolded.

which the controller was distracted when a controller in the next sector asked him to descend an aircraft that he had just climbed to a higher altitude. After that time, the controller seemed to be less active, letting events happen instead of managing them effectively. Thus, this factor was called "Inactivity."

Component 2 included the items Operational Deviations/Special Use Airspace Violations, Incorrect Information in the Computer, and (to a lesser extent) Transmission Errors and Operational Errors. In some sectors handling St. Louis departures, controllers sometimes failed to change altitude information already displayed in a data block or entered new altitudes in the data block that did not match the altitude clearance they had given the pilot. The reason for this was that the altitude limits for these sectors prevented controllers from clearing pilots above certain altitudes, but there was a certain amount of workload associated with entering interim altitudes for departures and changing them later. In some traffic samples, the controllers failed to update some (but not all) data blocks so as to reduce their workload. Many of the SME participants evaluated these actions as errors. Because of their failure to systematically update altitude information in aircraft data blocks, this component was called "Disorganization."

Component 3 included the items Unnecessary Delays and Transmission Errors. Operational Errors were negatively correlated with this factor and Failed to Accommodate Pilot Requests had a small positive correlation. During the traffic samples, unnecessary delays often involved failing to clear departing aircraft to higher altitudes in a timely way, failing to allow pilots to proceed to a higher altitude when requested and failing to clear pilots to go direct as requested. Most of the delays and transmission errors occurred during one traffic sample in which the (very busy) controller frequently asked pilots to repeat what they had said. Although these delays and transmission errors resulted from the controller's difficulty in effectively keeping up with the traffic situation, the controllers continued to maintain separation between aircraft. Thus, this component was called "Inefficient but Safe."

Relationships among reduced variable set. The previous analyses identified a reduced set of variables describing controller workload and performance. Intercorrelations among these variables are shown in Table 9.

By definition, a Varimax rotation produces orthogonal components, so correlations among the three BEC components were 0, as was the correlation between the two Workload components. The Average OTS Rating was significantly correlated with all other measures, both performance and workload. Correlations with the BEC component scores (based on errors recorded by the participants) and with the (reverse scored) TLX Performance scale were both

Table 9. *Correlations Among Reduced Set of Performance and Workload Measures (N=128).*

	Avg OTS rating	BEC1: Inactivity	BEC2: D-org	BEC3: Ineff-S	TLX Perf	Wkld1: Activity	Wkld2: Frustration
Average OTS rating	1.0						
BEC1: Inactivity	-.25**	1.0					
BEC2: Disorganization	-.27**	0	1.0				
BEC3: Inefficient but Safe	-.21*	0	0	1.0			
TLX Performance	-.52**	.26**	.29**	.17	1.0		
Wkld1: Activity	.21*	.11	.11	.17	-.17	1.0	
Wkld2: Frustration	.19*	. 24**	-.04	-.10	.14	0.0	1.0

Note: * $p < .05$; ** $p < .01$; **Abbreviations:** TLX Perf = TLX Performance, Inactivity = Inactivity component, D-org = Disorganization component, Ineff-S = Inefficient but Safe component.

negative. The TLX Performance scale was also significantly correlated with the BEC Inactivity and Disorganization components (but not with the BEC Inefficient but Safe component). BEC Inactivity had a significant positive correlation with the Workload Frustration component (but not with the Workload Activity component, as might have been expected). Besides the significant correlations with OTS and the BEC Inactivity components, the two workload components were not significantly correlated with any other variables.

Summary of Part 1 Results

Twenty-four controller performance and mental workload variables were obtained from SMEs who observed eight 20-minute traffic samples. To simplify later analyses, these variables were combined into seven composite controller performance and mental workload measures. Five composite controller performance measures were derived from the OTS rating form, the BEC, and the NASA TLX Performance scale. These were the Average OTS rating, the TLX Performance scale, and three principal components derived from the BEC: 1) Inactivity, 2) Disorganization, and 3) Inefficient but Safe. Two composite mental workload measures were derived from five NASA TLX scales, the ATWIT on-line workload ratings, and SME activity level ratings. The two resulting mental workload principal components were called Activity and Frustration.

Part 2: Assessment of Validity of POWER Measures

At this point, the level of analysis becomes the traffic sample and not the observer. Values for the POWER measures were calculated for each traffic sample. All the performance measures and all but the ATWIT mental workload measure were provided only once for each traffic sample. Static complexity (based on sector characteristics) did not often vary when the same sector was observed on two occasions (except when a sector was split out during one traffic sample and combined in the other). On the other hand, dynamic complexity did vary across traffic samples. Because many of the variables in this analysis were measured only once per traffic sample, the following analysis was conducted with an N of eight traffic samples.

Values for the seven controller performance and mental workload measures were averaged across raters for each traffic sample. The static, dynamic, and overall complexity factors and the POWER measures were computed for each traffic sample. Descriptive statistics, averaged across traffic sample, are shown in Table 10. Some of the POWER measures (primarily certain kinds of data entries, such as handoffs and altitude changes) occurred fairly often, on the average, over the 20-minute periods. Other data entries (e.g., pointouts, data block offsets, distance reference indicators [DRIs, also known as J-rings], track reroutes, and strip requests) did not occur very often (less than once every 20 minutes). The complexity measures, which were standardized, and the controller performance and mental workload measures, derived from orthogonally-rotated principal components, had mean values of zero, but the standard deviations indicate their relative variability.

Tables 11-13 show correlations of the POWER measures with the sector complexity, controller performance, and mental workload measures, respectively. Correlations significant at the .05 level or lower are indicated by **. Since the number of traffic samples analyzed was so small (N=8) and the number of correlations computed was so large (N=260), it is likely that many of the statistically significant correlations occurred due to chance. However, this result is less likely if a POWER measure was correlated with more than one measure of a construct or if several similar POWER measures were correlated with the same construct. On the other hand, many of the constructs are independent (because they are component scores produced by a principal components analysis with varimax rotation). Thus, a lack of relationship of a POWER measure with multiple components is not unexpected.

Relationship with Sector Complexity. Table 11 shows the relationship of POWER measures with the three measures of sector complexity. The sector complexity measures were related to several POWER measures. Higher static complexity (based on sector characteristics) was related to higher average speed variation, fewer R controller pointouts, and fewer data block offsets. Higher dynamic complexity (based on situational characteristics) was related to longer times that aircraft were under control. Higher overall complexity (combining the components of both static and dynamic complexity) had no significant correlations with any of the POWER measures.

Table 10. *Descriptive Statistics for POWER, Sector Complexity, Controller Performance, and Mental Workload Measures Averaged over Traffic Samples (N=8).*

	Descriptive Statistics	
Power Measures	Mean	SD
Total N aircraft controlled	15.25	5.23
Max aircraft controlled simultaneously	6.88	2.47
Average time aircraft under control	389.75	97.62
Avg Heading variation	11.64	3.02
Avg Speed variation	1.28	.68
Avg Altitude variation	.84	.51
Total N altitude changes	12.13	5.38
Total N handoffs	19.50	7.37
Total N handoffs accepted	5.88	3.98
Avg time to accept handoff	39.22	19.54
Total N handoffs initiated	10.00	4.00
Avg time until initiated HOs are accepted	50.47	27.26
N Radar controller data entries	56.75	22.70
N Radar controller data entry errors	1.13	1.36
N Data controller data entries	9.63	5.73
N Data controller data entry errors	0.38	0.52
N Route displays	2.00	2.27
N Radar controller pointouts	0.38	0.74
N Data controller pointouts	0.38	1.06
N data block offsets	0.75	0.89
Total N Conflict Alerts displayed	0.50	0.53
Number of Conflict Alert suppression entries	0.13	0.35
N Distance Reference Indicators requested	0.25	0.46
N Distance Reference Indicators deleted	0.13	0.35
N track reroutes	0.38	0.74
N strip requests	0.13	0.35
Complexity measures		
Static complexity	0.0	2.31
Dynamic complexity	0.0	5.57
Overall complexity	0.0	7.62
Performance Measures		
Average OTS Rating	3.79	0.24
TLX Performance	44.53	4.89
BEC Inactivity component	0.0	0.47
BEC Disorganization component	0.0	0.13
BEC Inefficient but safe component	0.0	0.33
Workload Measures		
Workload activity component	0.0	0.59
Workload frustration component	0.0	0.52

Table 11. *Correlations of POWER Measures with Measures of Sector complexity (N=8).*

Power Measure	Complexity Measure		
	Static complexity	Dynamic complexity	Overall complexity
Total N aircraft controlled	-.52	.16	-.20
Max aircraft controlled simultaneously	-.39	.66	.25
Average time aircraft under control	-.14	.72**	.44
Avg Heading variation	.58	-.40	.06
Avg Speed variation	.72**	.18	.57
Avg Altitude variation	.38	.18	.36
Total N altitude changes	.25	.37	.42
Total N handoffs	-.51	-.13	-.41
Total N handoffs accepted	-.61	.57	.05
Avg time to accept handoff	-.41	-.57	-.67
Total N handoffs initiated	-.52	-.17	-.44
Avg time until initiated HOs are accepted	.62	.08	.44
N Radar controller data entries	-.29	.48	.17
N Radar controller data entry errors	.12	.48	.43
N Data controller data entries	-.20	-.17	-.25
N Data controller data entry errors	.32	-.34	-.06
N Route displays	-.35	-.02	-.22
N Radar controller pointouts	-.78**	.33	-.23
N Data controller pointouts	-.66	.00	-.40
N data block offsets	-.81**	.36	-.23
Total N Conflict Alerts Displayed	.20	.49	.48
Number of Conflict Alert suppression entries	.62	.34	.63
N Distance Reference Indicators requested	-.29	.40	.11
N Distance Reference Indicators deleted	-.05	-.17	-.15
N track reroutes	.57	.24	.52
N strip requests	.46	.27	.48

Note: ** $p < .05$

Relationship with controller performance. Table 12 shows the relationship of POWER measures with measures of controller performance. All the controller performance measures were related to several POWER measures. Higher Average OTS ratings were related to more Conflict Alerts displayed. Higher TLX performance scores (indicating lower performance) were related to more R and D controller pointouts made, and more data block offsets. It was also related to lower average heading variation. Higher scores on the BEC Inactivity scale were related to lower average heading variation, more handoffs accepted, and more R and D controller pointouts. Higher scores on the BEC Disorganization scale were related to higher altitude variation and more DRIs deleted. Higher scores on the BEC Inefficient but Safe scale were related to more aircraft controlled simultaneously, more handoffs accepted, more R and D controller pointouts, and more data block offsets.

Relationship with Mental Workload. Table 13 shows the relationship of POWER measures with measures of mental workload. Higher scores on the Workload

Table 12. *Correlations of POWER Measures with Measures of Controller Performance (N=8).*

	Controller Performance Measure				
Power Measure	Average OTS Rating	TLX Performance	BEC Inactivity	BEC Disorganization	BEC Inefficient but safe
Total N aircraft controlled	.36	.23	.46	-.43	.62
Max aircraft controlled simultaneously	.43	.26	.56	-.30	.78**
Average time aircraft under control	-.05	.29	.47	.40	.60
Avg Heading variation	.45	-.91**	-.71**	-.11	-.57
Avg Speed variation	-.11	-.38	-.29	.51	-.12
Avg Altitude variation	-.09	-.24	-.25	.72**	-.18
Total N altitude changes	.14	-.26	.04	.51	.38
Total N handoffs	.42	.09	.25	-.42	.35
Total N handoffs accepted	-.02	.65	.79**	-.01	.91**
Avg time to accept handoff	-.01	-.05	-.12	-.20	-.22
Total N handoffs initiated	.42	.11	.24	-.44	.29
Avg time until initiated HOs are accepted	-.01	-.18	-.42	-.06	-.38
N Radar controller data entries	.66	-.08	.21	-.35	.44
N Radar controller data entry errors	-.00	.25	.08	-.42	.02
N Data controller data entries	.07	.20	-.13	.18	.01
N Data controller data entry errors	.44	-.40	-.51	-.69	-.53
N Route displays	.33	.18	-.16	.20	-.04
N Radar controller pointouts	-.34	.80**	.95**	-.15	.89**
N Data controller pointouts	-.56	.74**	.88**	-.10	.79**
N data block offsets	-.04	.75**	.70	-.01	.72**
Total N Conflict Alerts Displayed	.83**	-.39	-.07	-.24	.24
Number of Conflict Alert suppression entries	.26	-.58	-.10	.16	.21
N Distance Reference Indicators requested	.05	.30	-.01	.51	.08
N Distance Reference Indicators deleted	-.35	.19	-.24	.79**	-.21
N track reroutes	.08	-.46	-.21	.52	.10
N strip requests	-.01	.00	-.09	-.47	-.17

Note: ** $p < .05$

Activity scale were related to more R controller data entries and more Conflict Alerts displayed. Higher scores on the Workload Frustration scale were related to more aircraft controlled, more handoffs accepted, more R and D controller pointouts, and more data block offsets.

Analysis of data obtained at four-minute increments. One final analysis was conducted that utilized POWER measures and ATWIT ratings obtained at four-minute intervals for the same traffic samples. The purpose of the analysis was to compare POWER measures with an indicator of the workload activity component using more observations. Because the POWER measures could be computed and the ATWIT ratings obtained at four-minute intervals, 40 observations were available for analysis.

Table 14 shows means and standard deviations for the POWER measures and ATWIT ratings obtained at 4-minute intervals. Comparing the statistics for the POWER measures in Table 14 with those listed in Table 10 shows that the means and standard deviations in this table are lower than they were in

Table 13. *Correlations of POWER Measures with Mental Workload Measures (N=8).*

	Mental Workload Measure	
Power Measure	Workload Activity	Workload Frustration
Total N aircraft controlled	.56	.73**
Max aircraft controlled simultaneously	.69	.64
Average time aircraft under control	.34	.28
Avg Heading variation	.31	-.43
Avg Speed variation	-.04	-.30
Avg Altitude variation	.14	-.44
Total N altitude changes	.53	.29
Total N handoffs	.58	.51
Total N handoffs accepted	.33	.81**
Avg time to accept handoff	-.08	.15
Total N handoffs initiated	.57	.41
Avg time until initiated HOs are accepted	-.45	-.40
N Radar controller data entries	.83**	.38
N Radar controller data entry errors	-.44	-.07
N Data controller data entries	.05	.23
N Data controller data entry errors	-.08	-.36
N Route displays	.40	.01
N Radar controller pointouts	-.04	.91**
N Data controller pointouts	-.28	.90**
N data block offsets	.24	.72**
Total N Conflict Alerts Displayed	.96**	.03
Number of Conflict Alert suppression entries	.47	.06
N Distance Reference Indicators requested	.16	.00
N Distance Reference Indicators deleted	-.26	-.11
N track reroutes	.32	.00
N strip requests	-.50	-.30

Note:** $p < .05$

Table 10. Reducing the time period analyzed reduces the number of aircraft available for analysis and allows only shorter time segments to be considered.

Table 15 shows the correlations of the POWER measures computed at 4-minute intervals with ATWIT ratings obtained for the same time periods. This analysis produced more significant correlations than did the previous analysis based on only eight observations. The results were similar (but not identical) to the correlations based upon an N of 8 between the POWER measures and the Workload Activity scale. In this analysis, higher ATWIT ratings were related to more aircraft controlled and controlled simultaneously, more altitude changes made, more total handoffs made, more R controller data entries (but not D entries), and more Conflict Alerts displayed.

Summary of Part 2 Results

The mental workload, controller performance, and sector complexity measures and factor scores described in Part 1 of this paper were correlated with the POWER measures to assess the validity of the POWER measures in predicting controller workload and performance. Analyses were initially conducted using the eight traffic samples as observations. All of the measures had significant correlations with more than one POWER measure. An additional analysis was conducted that correlated the POWER measures computed for four-minute intervals with the ATWIT ratings obtained at the same rate. This analysis yielded a greater number of significant correlations than did the one based on the eight traffic samples.

Table 14. *Descriptive Statistics for POWER, Sector Complexity, Controller Performance, and Mental Workload Measures Obtained for4-Minute Periods Averaged over Traffic Samples (N=40).*

	Descriptive Statistics	
Power Measures	Mean	SD
Total N aircraft controlled	7.20	2.73
Max aircraft controlled simultaneously	5.48	2.35
Average time aircraft under control	158.35	34.38
Avg Heading variation	1.13	0.88
Avg Speed variation	4.33	2.51
Avg Altitude variation	2.07	1.52
Total N altitude changes	3.50	2.20
Total N handoffs	3.85	2.02
Total N handoffs accepted	1.15	1.12
Avg time to accept handoff	25.91	27.58
Total N handoffs initiated	1.98	1.29
Avg time until initiated HOs are accepted	41.00	45.45
N Radar controller data entries	11.35	5.54
N Radar controller data entry errors	0.23	0.58
N Data controller data entries	1.93	2.04
N Data controller data entry errors	0.08	0.27
N Route displays	0.40	0.84
N Radar controller pointouts	0.08	0.27
N Data controller pointouts	0.08	0.47
N data block offsets	0.15	0.43
Total N Conflict Alerts displayed	0.08	0.27
Number of Conflict Alert suppression entries	0.05	0.22
N Distance Reference Indicators requested	0.05	0.22
N Distance Reference Indicators deleted	0.03	0.16
N track reroutes	0.08	0.27
N strip requests	0.03	0.16
ATWIT rating	2.76	0.59

Discussion and Conclusions

Twenty-four measures of controller workload and performance were collected from 16 SME observers to assess the validity of a set of taskload measures derived independently from routinely recorded ATC data. However, before conducting the validity analysis, it was necessary to analyze the SME measures to determine their reliability, assess their inter-relationships, and determine whether a smaller set of measures could be identified to replace the larger set in the later validity analysis.

The results of the analyses described here suggest that most of the measures were reliable. One exception was the OTS rating scale. When this scale was used previously (as described in Bruskiewicz et al., 2000), the raters had undergone a joint two-week training session to ensure their reliability. The participants in this study were not able to undergo an equivalent amount of training on the use of the scale because they did not have sufficient time available (as is usually true for SMEs who participate in human factors research studies). The reduced reliability confirms the requirement for extensive rater training (as described in Sollenberger, Stein, & Gromelski, 1997).

Table 15. *Correlations of POWER Measures with ATWIT Ratings (N=40).*

Power Measure	ATWIT Rating
Total N aircraft controlled	.80**
Max aircraft controlled simultaneously	.77**
Average time aircraft under control	.40
Avg Heading variation	.12
Avg Speed variation	-.06
Avg Altitude variation	.10
Total N altitude changes	.43**
Total N handoffs	.47**
Total N handoffs accepted	.40
Avg time to accept handoff	.15
Total N handoffs initiated	.36
Avg time until initiated HOs are accepted	.01
N Radar controller data entries	.65**
N Radar controller data entry errors	-.02
N Data controller data entries	-.07
N Data controller data entry errors	-.04
N Route displays	.10
N Radar controller pointouts	.17
N Data controller pointouts	.01
N data block offsets	.16
Total N Conflict Alerts Displayed	.44**
Number of Conflict Alert suppression entries	.35
N Distance Reference Indicators requested	.11
N Distance Reference Indicators deleted	-.11
N track reroutes	.06
N strip requests	-.30

Note: ** $p < .01$

Reduced reliability using the OTS rating scale may also have been affected by the lack of available information. For example, D-side communications were not available, the participants were unable to determine some of the display settings used by the controllers, and the participants only observed recorded data, not the actual ATC situation.

These analyses also suggest that it is possible to identify a reduced set of variables that adequately describe mental workload and controller performance. The set identified here consisted of five composite controller performance measures: The Average OTS rating, the TLX Performance scale, and three principal components derived from the BEC (Inactivity, Disorganization, and Inefficient but Safe); and two composite mental workload measures (Activity and Frustration). While it is possible to identify a reduced set of measures, they may not all be equally effective. For example, using the OTS rating scale requires extensive rater training, which was apparently not accomplished as successfully in this study as it was in the AT-SAT High Fidelity simulation study. The use of the NASA TLX to measure mental workload is also somewhat questionable-four of the scales were so highly correlated that they seemed to be measuring the same construct, while another (Performance) was completely unrelated to the remaining workload scales.

It appears that asking SMEs to observe recorded ATC traffic samples instead of live ATC activity may be a reasonable way to obtain assessments of mental workload and controller performance. Subsequent research should investigate whether obtaining additional information about the sectors, procedures, and activities (such as recordings of D controller communications), could enhance the observers' understanding of the traffic samples. If additional information can be obtained, it may be possible to enhance the OTS rating process and obtain more reliable and valid ratings of controller effectiveness. Increasing the number of times the measures are obtained and/or the number of traffic samples observed would increase the number of observations available for later analysis.

Using this set of reduced measures, a correlational analysis was performed. However, given the small number of observations, the results should be interpreted with caution. Nevertheless, some interesting relationships between controller and sector activities and the constructs of sector complexity, controller performance, and mental workload were evident.

The interpretation of the relationship between POWER measures and the sector complexity measures may provide an explanation for the different aspects of sector complexity. Certain POWER measures may have been related to static complexity (based on sector characteristics) because of the structure of the sectors included in the study. For example, higher speed variation (suggesting more speed changes were issued-probably because the sectors were arrival or departure sectors), making fewer data block offsets, and fewer R controller pointouts during traffic samples may be related to the way the sector was configured. Moreover, other POWER measures may also have been related to dynamic complexity (based on events that occurred during the traffic sample) because of the structure and function of the sector. For example, controlling more aircraft simultaneously and having aircraft under control for a longer time are related to sector size, busyness, and purpose (i.e., arrival, departure, overflight sector).

The relationships between the POWER measures and measures of controller performance are not as easy to interpret. Specifically, higher performance ratings (on the reverse-scaled TLX Performance scale) and higher inactivity were related to lower average heading variation (suggesting controllers who received poorer performance ratings turned or vectored aircraft less often), more handoffs accepted, more R and D controller pointouts, and more data block offsets. Likewise, accepting more handoffs, making more pointouts and more data block offsets were also related to higher inefficiency. Perhaps the controllers making these entries were more engaged in housekeeping activities than efficiently and effectively handling the traffic. Higher altitude variation (suggesting more clearances involving altitude changes were issued) and more DRIs deleted were related to higher disorganization scores. Finally, more conflict alerts displayed were related to higher average OTS ratings. This last result suggests that the OTS performance ratings made by the SME observers may have been partially based on the workload the participant perceived to occur during the traffic sample.

Mental workload seems to be related to certain aircraft and controller activities. The two workload components used in this study (activity and frustration) appear to measure different aspects of workload, in part because they were typically correlated with different POWER measures. Activity was related to R controller data entries and Conflict Alerts displayed (which was also significantly correlated with the Average OTS rating). Data entries and conflict alerts are an indicator of how busy the controller is. On the other hand, Frustration was related to total numbers of aircraft controlled, total handoffs accepted, R and D controller pointouts, and data block offsets made. This component seems to be related to the extent to which higher aircraft activity requires additional controller effort (such as pointing out an aircraft to another sector or moving data blocks to be able to continue to see aircraft information).

Part of the data were reanalyzed by comparing the ATWIT ratings, obtained in more but smaller increments, with POWER measures computed over a shorter period of time. This analysis found that, in addition to the variables related to the Activity and Frustration components in the other correlational analysis, higher ATWIT ratings were also related to the maximum number of aircraft controlled simultaneously and the numbers of altitude changes made.

Several POWER measures appear to be unrelated to any measure of sector complexity, controller performance, or mental workload. These included average time until initiated handoffs are accepted (hypothesized to be related to mental workload), number of R controller data entry errors (hypothesized to be related to controller performance), number of D controller data entries (hypothesized to be related to mental workload), number of route displays (hypothesized to be related to controller performance), number of distance reference indicators requested (hypothesized to be related to both controller performance and mental workload), number of track reroutes (hypothesized to be related to mental workload) and number of strip requests (hypothesized to be related to mental workload). While some of the hypothesized relationships may not exist, some of these variables occurred infrequently in this limited number of traffic samples, so it may be inappropriate to conclude, based on this study, that they are not at all related to any of the constructs.

Also, the Overall Complexity construct was not related to any POWER measures. This result suggests that it may be useful to distinguish between static and dynamic complexity rather than combining their influence into a single variable.

While this exploratory study has provided important information about the POWER measures, additional research is needed to better understand the relationships observed here. It may be possible to compare controller performance and mental workload ratings collected during simulation studies with POWER measures obtained for those traffic samples to obtain additional evidence about the validity of the measures.

It will also be necessary to analyze larger blocks of POWER data to examine the statistical characteristics and interrelationships of the measures, and perhaps identify a smaller set of POWER measures that account for differences in sector complexity, controller performance, and mental workload. When the properties and limitations of these measures are better understood, they may then be used to calculate baseline measures for the current National Airspace System and may eventually be used to assess the effects of implementing new ATC systems.

References

Borman, W.C., Hedge, J.W., Hanson, M.A., Bruskiewicz, K.T., Mogilka, H., Manning, C., Bunch, L.B., & Horgen, K.E. (2001). Development of criterion measures of air traffic controller performance. In R.A. Ramos, M.C. Heil, & C.A. Manning (Eds.), Documentation of validity for the AT-SAT computerized test battery: Volume II. (Report No. DOT/FAA/AM-01/6). Washington, DC: FAA Office of Aviation Medicine.

Bruskiewicz, K.T., Hedge, J.W., Manning, C.A., & Mogilka, H.J. (2000, January). The development of a high fidelity performance measure for air traffic controllers. In C.A. Manning (Ed.), Measuring air traffic controller performance in a high-fidelity simulation. (Report No. DOT/FAA/AM-00/2). Washington, DC: FAA Office of Aviation Medicine.

Buckley, E.P., & Beebe, T. (1972). The development of a motion picture measurement instrument for aptitude for air traffic control (Report No. FAA-RD-71-106). Washington, DC: Federal Aviation Administration, Systems Research and Development Service.

Buckley, E.P., DeBaryshe, B.D., Hitchner, N., & Kohn, P. (1983). Methods and measurements in real-time air traffic control system simulation (Report No. DOT/FAA/CT-83/26). Atlantic City, NJ: DOT/FAA Technical Center.

Dailey, J.T. & Pickrel, E.W. (1984). Development of the Multiplex Controller Aptitude Test. In S.B. Sells, J.T. Dailey, and E.W. Pickrel, (Eds.), Selection of air traffic controllers. (Report No. DOT/FAA/AM-84/2). Washington, DC: FAA Office of Aviation Medicine.

Federal Aviation Administration. (1991). Multiple Virtual Storage (MVS); Subprogram Design Document; National Track Analysis Program (NTAP). (Report No. NASP-9114-H04). Washington, DC: Author.

Federal Aviation Administration. (1993). Multiple Virtual Storage (MVS); User's Manual; Data Analysis and Reduction Tool (DART). (Report No. NASP-9247-PO2). Washington, DC: Author.

Federal Aviation Administration. (2000). Air Traffic Control. (Order 7110.65M). Washington, DC: Author.

Galushka, J., Frederick, J., Mogford, R., & Krois, P. (1995, September). Plan View Display baseline research report. (Report No. DOT/FAA/CT-TN95/45). Atlantic City, NJ: Federal Aviation Administration Technical Center.

Grossberg, M. (1989, April). Relation of airspace complexity to operational errors. Quarterly Report of the Federal Aviation Administration's Office of Air Traffic Evaluation and Analysis.

Hadley, G.A. Guttman, J.A., & Stringer, P.G. (1999, June). Air traffic control specialist performance measurement database. (Report No. DOT/FAA/CT-TN99/17). Atlantic City, NJ: William J. Hughes Technical Center.

Hanson, M.A., Borman, W.C., Mogilka, H.J., Manning, C.A., & Hedge, J.W. (1999). Computerized assessment of skill for a highly technical job. In F. Drasgow, & J.B. Olson-Buchanan, (Eds.) Innovations in computerized assessment. Mahwah, NJ: Lawrence Erlbaum Associates.

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock and N. Meshkati (Eds.), Human mental workload (pp. 139-183). Amsterdam: North-Holland.

Mills, S.H., Manning, C.A., & Pfleiderer, E M. (1999, May). Computing en route baseline measures with POWER. Poster presented at Tenth International Symposium on Aviation Psychology, Columbus, OH.

Mogford, R.H., Murphy, E.D., Roske-Hofstrand, R.J., Yastrop, G., & Guttman, J.A. (1994, June). Research techniques for documenting cognitive processes in air traffic control: Sector complexity and decision making (Report No. DOT/FAA/CT-TN94/3). Atlantic City, NJ: Federal Aviation Administration Technical Center.

Moroney, W.F., Biers, D.W., and Eggemeier, F.T. (1995). Some measurement and methodological considerations in the application of subjective workload measurement techniques. The International Journal of Aviation Psychology, 5, 87-106.

Nygren, T.E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. Human Factors, 33, 17-33.

Ramos, R.A., Heil, M.C., & Manning, C.A. (Eds.), (2001). Documentation of validity for the AT-SAT computerized test battery: Volumes I and II. (Report No. DOT/FAA/AM-01/6). Washington, DC: FAA Office of Aviation Medicine.

Rodgers, M.D., & Duke, D.A. (1993). SATORI: Situation Assessment Through Re-creation of Incidents. The Journal of Air Traffic Control, 35(4), 10-14.

Rodgers, M.D., Mogford, R.H., & Mogford, L.S. (1998). The relationship of sector characteristics to operational errors. (Report No. DOT/FAA/AM-98-14). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Sollenberger, R.L., Stein, E.S., & Gromelski, S. (1997). The development and evaluation of a behaviorally-based rating form for assessing air traffic controller performance. (Report No. DOT/FAA/CT-TN96-16). Atlantic City, NJ: Federal Aviation Administration Technical Center.

Stein, E.S. (1985). Air traffic controller workload: An examination of workload probe. (Report No. DOT/FAA/CT-TN84/24). Atlantic City, NJ: Federal Aviation Administration Technical Center.

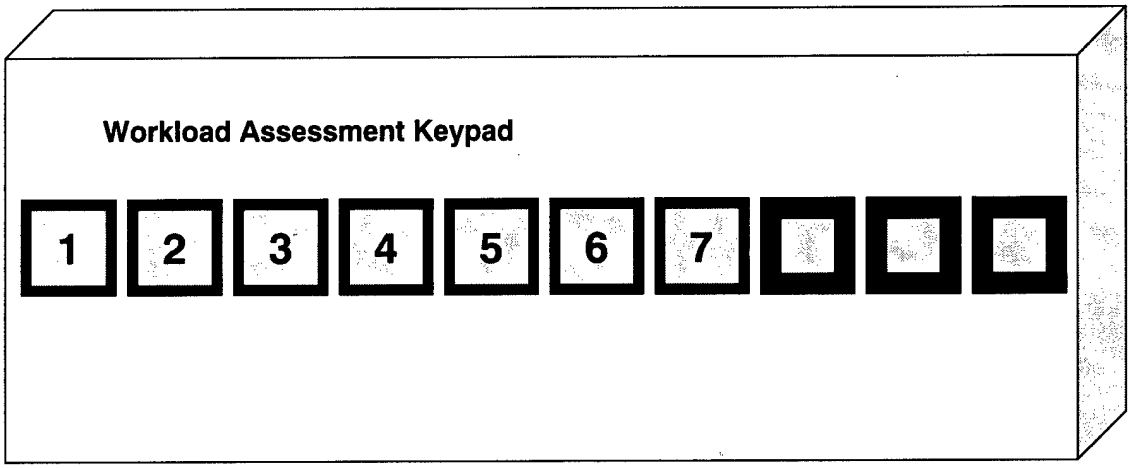
Tucker, J.A. (1984). Development of dynamic paper-and-pencil simulations for measurement of air traffic controller proficiency. In S.B. Sells, J.T. Dailey & E.W. Pickrel (Eds.), Selection of air traffic controllers (Report No. FAA-AM-84-2, pp. 215-241). Washington, DC: FAA Office of Aviation Medicine.

Wickens, C.D., Mavor, A.S., Parasuraman, R., & McGee, J.P. (Eds). (1998). The future of air traffic control: Human operators and automation (pp. 216-217). Washington, DC: National Academy Press.

Wyndemere Inc. (1996). An evaluation of air traffic control complexity (Contract number NAS2-14284). Boulder, CO: Author.

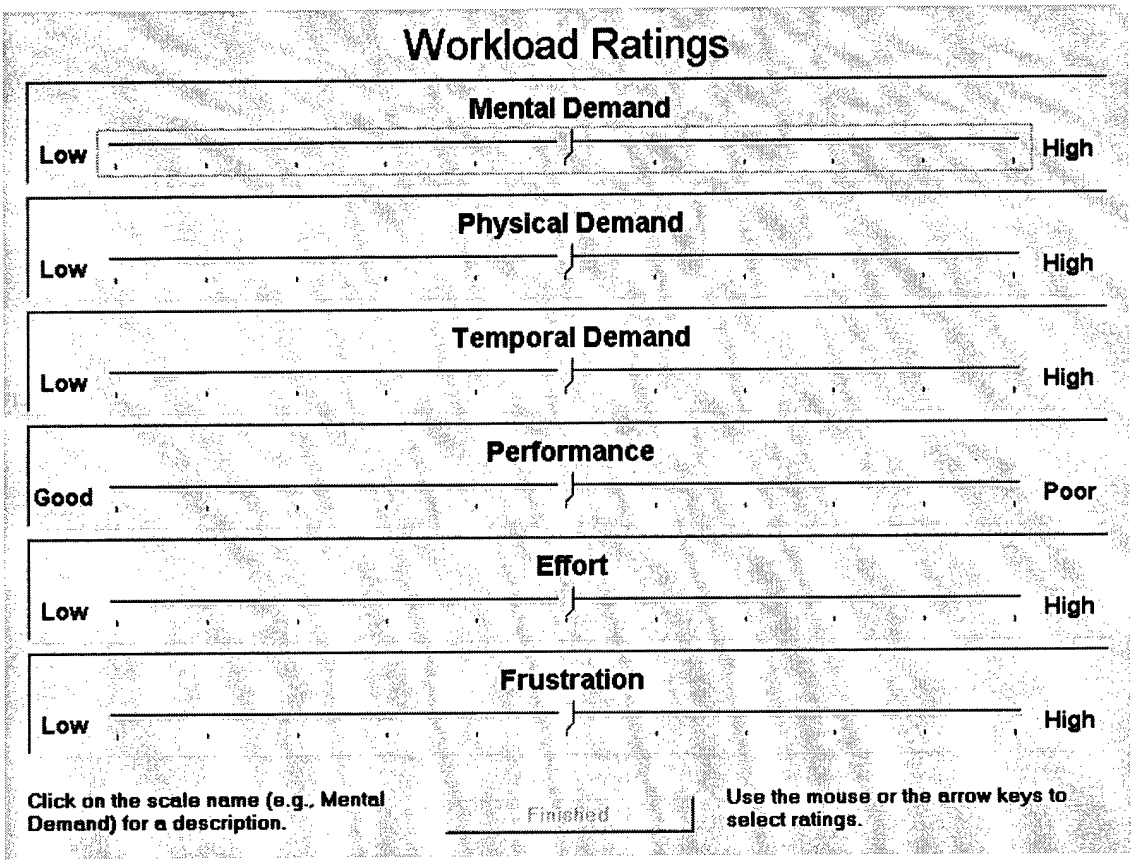
Appendix A

Diagram of Workload Assessment Keypad (WAK) used to enter Air Traffic Workload Input Technique (ATWIT) ratings



Appendix B

Computerized screen used to enter TLX workload ratings



Appendix C

Instructions for completing computerized version of NASA TLX

In this study, you will observe re-creations of air traffic activity controlled by other controllers. We are interested in finding out your perception of how difficult you thought his or her task was and how well you thought the person performed the task. Our objective is to measure your perception of their "workload" level. The concept of workload is composed of several different factors. Therefore, we would like you to tell us about several individual factors rather than one overall workload score.

Here is an example of the rating scales. As you can see, there are six scales on which you will be asked to provide a rating score: *mental demand, physical demand, temporal demand, effort, frustration, and performance.*

Rating Scales

Mental demand refers to the level of mental activity like thinking, deciding, and looking that was required to perform the task. You will rate this scale from low to high.

Physical demand involves the amount of physical activity required of the controller, such as controlling or activating.

Temporal demand refers to the time pressure you think the controller experienced during the task. In other words, was the pace slow and leisurely or rapid and frantic? If the pace was rapid and frantic then he or she experienced high temporal demand.

Effort refers to how hard you think the controller worked (both mentally and physically) in order to achieve his or her level of performance.

Frustration level refers to how secure and relaxed versus stressed and discouraged you think the controller felt during the task. If you think he or she felt secure and relaxed, then you should provide a rating of low frustration.

Performance level refers to your perception of the controller's performance level. Your rating here should reflect your satisfaction with his or her performance in accomplishing the goals of the task.

Making your response

You should indicate your rating by adjusting the slider on the bar associated with each item. For example, if you want to give a high rating of stress factor, move the slider to the right of the half-way mark. The higher the stress rating, the closer the slider should be "HIGH." In contrast, if your stress rating is low, you would move the slider toward the "LOW" end of the line. Likewise, if the stress rating is average place the slider in the center of the line.

Appendix D

Traffic Sample Activity Level/Task Load Rating Scale

Please rate your impression of the activity level or task load of the traffic sample you just finished watching on the scale you see below. Please mark one of the 5 alternatives by making an X above the vertical line that extends above the description you think is appropriate.

Not at all busy	Slightly busy	Average busyness	Moderately busy	Very busy

Appendix E

POWER Validation Study Over The Shoulder (OTS) Rating Form
Administrative Information - Page 1

Participant ID Number:	Counterbalancing Order: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Sector: 14 30 52 54	Traffic Sample: A B T

POWER Validation Study Over The Shoulder (OTS) Rating Form

Rating Dimensions	Rating Scale: Not rated/ observed	Below Average	Fully Adequate	Exceptional
A. Maintaining Separation	(NA)	① ②	③ ④ ⑤	⑥ ⑦
• Checks separation and evaluates traffic movement to ensure separation standards are maintained				
• Considers aircraft performance parameters when issuing clearances				
• Detects and resolves impending conflictions				
• Establishes and maintains proper aircraft identification				
• Applies appropriate speed and altitude restrictions				
• Properly uses separation procedures to ensure safety				
• Analyzes pilot requests, plans and issues clearances				
• Issues safety and traffic alerts				
B. Maintaining Efficient Air Traffic Flow	(NA)	① ②	③ ④ ⑤	⑥ ⑦
• Accurately predicts sector traffic overload and takes appropriate action				
• When necessary, issues a new clearance to expedite traffic flow				
• Ensure clearances require minimum flight path changes				
• Reacts to/resolves potential conflictions efficiently				
• Controls traffic in a manner that ensures efficient and timely traffic flow				
C. Communicating Clearly, Accurately, and Efficiently	(NA)	① ②	③ ④ ⑤	⑥ ⑦
• Issues clearances that are complete, correct, and timely				
• Communicates clearly and concisely				
• Makes only necessary transmissions				
• Uses correct call signs				
• Uses standard/prescribed phraseology				
• Uses appropriate speech rate				
• Properly establishes, maintains, and terminates communications				
• Listens carefully to pilots and controllers				
• Avoids lengthy clearances				
• Issues appropriate arrival and departure information				
D. Technical Knowledge	(NA)	① ②	③ ④ ⑤	⑥ ⑦
• Is current on air traffic procedures.				
• Issues altitudes, speed adjustments, and turns within aircraft performance limitations				
• Is knowledgeable about ATC equipment and how it functions				
• Is knowledgeable about different aircraft capabilities				

AT-SAT High Fidelity Simulation Over The Shoulder (OTS) Rating Form - Page 2

Rating Dimensions	Rating Scale			
	Not rated/ observed	Below Average	Fully Adequate	Excep-tional
E. Prioritizing	(NA)	① ②	③ ④ ⑤	⑥ ⑦
• Performs air traffic duties needing immediate attention before less pressing ones	• Assesses potential air traffic problems that might result from own actions (e.g., revised clearances)			
• Uses appropriate priorities for control actions	• Takes early or prompt action to resolve air traffic problems			
• Accurately predicts problems that will result from circumstances such as rapidly degrading weather				
H. Overall Effectiveness		① ②	③ ④ ⑤	⑥ ⑦

Revised 8-13-99

Comments (optional):

Appendix F

Instructions for Over-the Shoulder (OTS) Rating Form

The Over-the-Shoulder (OTS) Rating Form will be completed after you watch a traffic sample. One form will be completed for each of the 4 practice and 8 actual traffic samples.

The OTS Rating Form contains five specific effectiveness categories and one overall effectiveness category that you will use to make assessment ratings as part of the POWER Validation Study. Each category includes a set of performance examples that describe the type of behaviors that should be considered when you assign your ratings.

To the right of the title for each effectiveness category, there is a 7 point rating scale. You will use this scale to evaluate the R-controller's effectiveness during the traffic sample. Your ratings should be based on your assessment of the R-controller's effectiveness in performing the behaviors listed under each category. Ratings of 1 or 2 indicate "Below Average" effectiveness. Ratings of 3, 4, or 5 indicate "Fully Adequate" effectiveness. Ratings of 6 or 7 indicate "Exceptional" effectiveness.

On the scales for the specific effectiveness categories, there is another point that you can mark labeled "NA". The name of this point is "Not rated/Observed." Please mark Not rated/Observed if you feel that watching the traffic sample did not provide enough information to allow you to make a rating for that specific effectiveness category. Notice that there is no "NA" category available to mark for the Overall Effectiveness Performance Rating.

When you finish watching a traffic sample, you will first complete the NASA TLX form located on the PC to the left of the SATORI workstation. Then you will total the errors you recorded on the BEC form. Finally, you will complete the OTS Rating Form.

Making Your Ratings

Read the performance examples listed under each specific effectiveness category. Then, compare your opinion about the controller's effectiveness during the traffic sample with the performance examples for that category.

After reviewing the performance examples for a specific effectiveness category, if you think the controller's effectiveness in that category was Below Average some of the time but was Fully Adequate more often, a rating of "3" would be best. Similarly, if you think a controller's effectiveness was Fully Adequate sometimes but was Exceptional more often, the fairest rating to give is probably a "6."

Once you have selected a rating, make your rating by blackening the appropriate circle on the OTS Rating Form. Again, if you feel that watching the traffic sample did not provide enough information to allow you to mark a specific effectiveness category, please mark "NA." However, even if you marked one or more "NAs" for a traffic sample, please fill in the Overall Effectiveness Rating.

Notes about completing OTS Rating Form:

- If you make a mistake when filling out the OTS form, erase the mark completely and fill in a different bubble.
- If you recorded no OEs on the BEC form, you may assign any number for the Maintaining Separation effectiveness category (A) on the OTS rating form. If you identified one OE, you should assign a rating no higher than 2 for Maintaining Separation. If you identified two OEs, you should assign a rating no higher than 1 for Maintaining Separation.
- When assigning the Overall Effectiveness rating, consider the controller's effectiveness in each of the specific effectiveness categories. Your Overall Effectiveness rating should be influenced most by the ratings you assigned to the specific effectiveness categories you think are most important. However, they should also be influenced to a lesser extent by the ratings you assigned to the specific effectiveness categories you think are less important.

Important Points to Remember when making OTS Ratings

- Try not to give a controller the same rating for all five specific effectiveness categories. Most people will perform well in some categories and less effectively in others. Your ratings should show the controller's strengths and weaknesses, as appropriate.
- Try not to give the same rating within each specific effectiveness category for all the traffic samples you observe. Instead, your ratings should indicate which controllers are performing more effectively and which are performing less effectively in each category.
- One thing to keep in mind is that the high effectiveness ratings (6 or 7) are truly outstanding. You should reserve these ratings, especially the "7," for the very highly effective controllers.
- If you know someone who controlled traffic in any of the traffic samples, please do not let that knowledge influence the ratings you assign.
- The *most* important point is to make your ratings as accurate as possible. This is the best way to help us validate the POWER measures.

Appendix G

POWER Validation Study

Behavior and Event Checklist

Event/Aircraft Identity		Totals
Operational Errors (Write all call signs in one box)	3.	
1.	4.	
2.	5.	
Operational Deviations/SUA violations (Write call signs in each box)	3.	
1.	4.	
2.	5.	

Behavior	Number of events	Totals
Failed to accept handoff		
LOA/Directive Violations		
Transmission errors		
Failed to accommodate pilot request		
Made late frequency change		
Unnecessary delays		
Incorrect information in computer		
Fail to issue weather information		

Participant ID #:	Counterbalancing Order: 1 2 3 4 5 6 7 8 / 9 10 11 12 13 14 15 16
Traffic Sample: A B T	Sector: 14 30 52 54

Appendix H

Instructions for Completing Behavior and Event Checklist (BEC)

The Behavior and Event Checklist (BEC) is used to record mistakes made by controllers or controller teams during the Traffic Samples you observe. The BEC was developed for the AT-SAT High Fidelity Simulation Exercise. You will complete one BEC form for each of the 4 "training" traffic samples and the 8 "experimental" traffic samples that you observe during this experiment.

You will record items on the BEC while the traffic sample is running. The first two types of events that you are to record are Operational Errors and Deviations. Note that in these traffic samples, no OEs or ODs were officially reported. Thus, it is not likely that you will observe one occur. However, if you think the controller or controller team you are watching committed an OE or OD, please record it on the BEC form. If you record an OE, please write the call signs of all involved aircraft in the same box. If you think an OD or SUA violation occurred, please write the call sign of the involved aircraft in one box.

When you are recording the other behaviors (those listed below the OEs and ODs), you need only make a tic mark in the box and do not need to record call signs. When you finish watching a traffic sample, you will first complete the NASA TLX form provided on the PC to the left of the SATORI workstation. Then you will total the errors recorded on the BEC form. Finally, you will complete the OTS Rating Form.

Notes about completing BEC form:

- If you make a mistake when filling out the BEC, either erase the mark or draw a squiggly line through the incorrect mark.
- The following list provides examples of special situations (other than when the standard rules would apply) when BEC items should be marked. This is not an exhaustive list.

Operational errors

- If an aircraft without Mode C doesn't report level, the controller doesn't determine a reported altitude, and the aircraft overflies another aircraft, it shall be scored as an OE. Also, if the controller doesn't enter a reported altitude in the computer, it shall also be scored as Incorrect Information in Computer.
- If an aircraft is cleared off an airport, but the controller is not yet talking to the aircraft, it is **NOT** an OE if another aircraft is cleared for approach into that same airport.

Operational Deviations

- An Operational Deviation is considered to occur if there is a violation of published MEAs or MIAs.
- An Operational Deviation is considered to occur if an aircraft comes within 2.5 miles of the airspace of another facility without being handed off or pointed out.
- An Operational Deviation occurred if the controller failed to point out an aircraft to the appropriate sector or if the controller issued a clearance to an aircraft while it is within another sector's airspace.

Special Use Airspace Violation

- A Special Use Airspace violation is considered to occur if an aircraft does not remain clear of an MOA or Restricted Area by either 3 NM or 500 feet of altitude. If an SUA violation occurs, it will be marked in the same area as Operational Deviations. Write call signs of involved aircraft in the boxes provided.

LOA/Directive Violation

- Violations of inter- and intra-facility LOAs will be considered LOA/Directive Violations.
- Count as LOA/Directive Violation if a frequency change is issued prior to completion of a hand-off for the appropriate aircraft.
- Count as LOA/Directive Violation if the controller makes a handoff to and switches the frequency to the incorrect facility.
- Count as LOA/Directive Violation if the controller drops a data block while the aircraft is still inside the airspace.
- Count as LOA/Directive Violation if the controller fails to inform the pilot of radar contact.
- Count as LOA/Directive Violation if the controller fails to coordinate inappropriate altitude for direction of flight within 2.5 miles of sector boundary.
- If controller fails to say "Radar service terminated," count as LOA/Directive violation and consider when making OTS ratings.

Transmission Errors

- Includes Readback/hearback errors
- Count as transmission error even if controller corrects himself/herself.

Failed to Accommodate Pilot Request

- A controller shall be rated as failing to accommodate a pilot request if he/she never takes appropriate action to accommodate the request, if the controller says unable when he/she could have accommodated the request, or if the controller says stand by and never gets back to the pilot. This situation applies if the rater determines that the controller could have accommodated the request without interfering with other activities.
- If another facility calls for a clearance and the controller fails to issue it unnecessarily, counts as Unnecessary Delay, not as Failure to Accommodate Pilot Request.

Made Late Frequency Change

- If an aircraft enters another sector without appropriate transfer of communications, the controller has made a Late Frequency Change.

Unnecessary Delay

- Includes accepting handoff late. Acceptance of handoff is considered late if the radar target is within 2.5 NM of 1) an Approach Control boundary if the aircraft is exiting Approach airspace or 2) crossing the sector boundary if the aircraft is transiting En-Route airspace.
- An Unnecessary Delay is considered to occur if a pilot request can be accommodated and the controller delays in doing so
- Count as Unnecessary Delay if the controller levels any departure at an altitude below the requested altitude and there was no traffic.
- Count as Unnecessary Delay if an aircraft in holding is not expeditiously cleared on course.
- If another facility calls for a clearance and the controller fails to issue it unnecessarily, counts as Unnecessary Delay, not as Failure to Accommodate Pilot Request.

Incorrect information in computer -

- Count incomplete or incorrect entries made by the R controller as Incorrect Information in Computer.
- If an aircraft does not have Mode C, the controller shall enter the reported altitude 1) when the pilot reports it, 2) prior to Handoff, or 3) by the end of the traffic sample. If this does not happen, count as Incorrect Information in Computer. Also, see OE.
- Altitude information in Data Blocks shall be considered incorrect if and when reported altitude differs by 1000 feet or more from assigned altitude displayed in same data block.
- Failure to correct within 2 minutes any incorrect entries made by D-controller that affect the R-side display is considered Incorrect Information in Computer.

Fail to Issue Weather Information

- Controllers must insure pilot has received current weather information. Issuance of enroute weather phenomenon is NOT required, but if issued should lead to a higher rating for doing so.